DOCUMENT RESUME

ED 219 457 TM 820 487

AUTHOR Jolly, S. Jean; And Others
TITLE Customization of the Achievement Testing Program for Palm Beach County Florida County Schools; Symposium.
PUB DATE Mar 82
NOTE 69p.; Papers presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, March 20-22, 1982).

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Achievement Tests; Criterion Referenced Tests; *Curriculum Evaluation; Educational Testing; Elementary Secondary Education; Evaluation Methods; Item Banks; Norm Referenced Tests; School Districts; *Test Construction; Testing Problems; *Testing Programs; *Test Validity
IDENTIFIERS *Palm Beach County Schools FL; Rasch Model; Stanford Achievement Tests

ABSTRACT

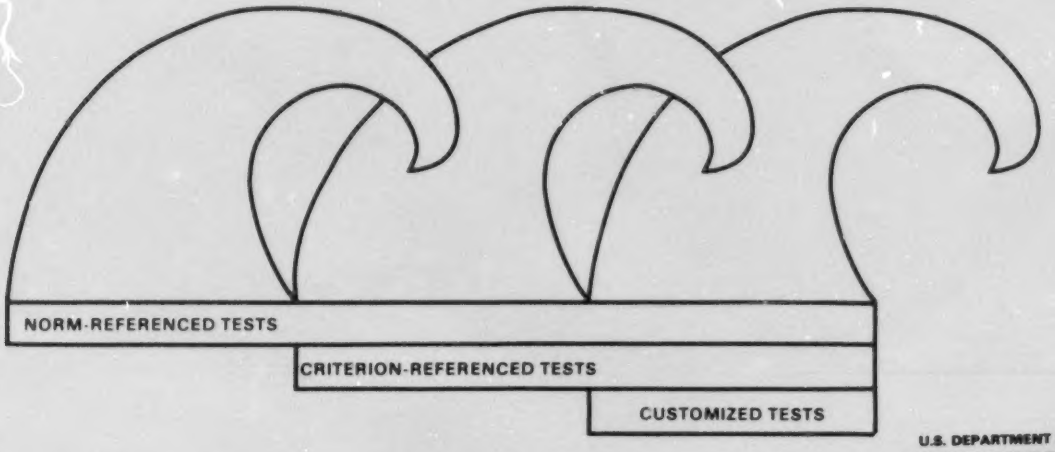
This document contains five papers describing the customization of the achievement testing program for Palm Beach County (Florida) Schools. They are as follows: "Customized Achievement Testing Programs" (S. Jean Jolly); "Using the Rasch Model to Develop Curriculum-Referenced Norms" (R. Robert Rentz); "Matching Standardized Achievement Test Items to Local Curriculum Objectives" (Ruth Bower); "An Exploratory Study of the Concept of Curriculum-Referenced Norms Using the Stanford Achievement Test, Sixth Edition" (Gary Gramenz); and "Supplementing Standardized Achievement Test Items for Testing of Local Objectives" (Bernard Jones). Inherent problems in standardized norm-referenced tests for curriculum evaluation are related to two sources of mismatch, content-not-taught and content-not-tested. A suggested solution is for the multiple use of standardized test items and the supplementation of selected standardized test items with items drawn from a locally-developed bank. Customization of the curriculum evaluation entails the derivation of a curriculum-referenced normative score and the linking of supplemental test items to selected standard items for criterion-referenced purposes. The computation of curriculum-referenced norms, comparison with a national sample on content known to be common, is facilitated if the Rasch Model has been the basis for test construction. Each reading and mathematics item on the Stanford Achievement Test, Seventh edition, to be administered in grades two through six, was examined for its match with local reading and mathematics objectives as a preliminary step in customizing an achievement testing program. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED219457

Symposium

CUSTOMIZATION OF THE ACHIEVEMENT TESTING PROGRAM FOR PALM BEACH COUNTY SCHOOLS



U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Presented at:

THE NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION

NEW YORK, NEW YORK
MARCH 20, 1982

TM 820 187

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

S. J. Jolly

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Session 3 C

CUSTOMIZATION OF THE ACHIEVEMENT TESTING PROGRAM FOR PALM BEACH COUNTY FLORIDA COUNTY SCHOOLS

Presented at:

National Council on Measurement in Education
New York, New York
March 20, 1982

PRESENTORS:

S. Jean Jolly
Director of Program Evaluation
Palm Beach County Florida Schools

Dr. R. Robert Rentz
Consultant
Psychological Corporation

Dr. Ruth Bower
Area Mathematics Consultant
Palm Beach County Florida Schools

Dr. Gary Gramenz
Specialist in Evaluation and Planning
Palm Beach County Florida Schools

Mr. Bernard Jones
Specialist in Test Development and Evaluation
Palm Beach County Florida Schools

DISCUSSANTS:

Dr. Roger Lennon
Briarcliffe Manor, New York

Dr. Howard Stoker
Florida State University

SESSION 3C

Customized Achievement Testing Programs
S. Jean Jolly

Using the Rasch Model to Develop Curriculum-References Norms
Dr. R. Robert Rentz

Matching Standardized Achievement Test Items to
Local Curriculum Objectives
Dr. Ruth Bower

An Exploratory Study of the Concept of Curriculum-Referenced Norms
Using the Stanford Achievement Test, Sixth Edition
Dr. Gary Gramenz

Supplementing Standardized Achievement Test Items
for Testing of Local Objectives
Mr. Bernard Jones

Appendix

CUSTOMIZED ACHIEVEMENT TESTING PROGRAMS

S. JEAN JOLLY

PALM BEACH COUNTY SCHOOLS

Presented at the

Annual Meeting of the

National Council on Measurement in Education

New York City

March, 1982

Historical Overview

Metaphor is a useful way of making connections between seemingly unconnected things. Alvin Toffler's (1981) wave metaphor is adopted here to view educational testing in its historical context. Much to the point in this symposium is Toffler's perception:

> Once I began thinking in terms of changes, colliding and overlapping, causing conflict and tension around us, it changed my perception of change itself....It became possible to distinguish those innovations that are merely cosmetic....from those that are truly revolutionary (p. 9).

We have had major changes in testing methodology, and these changes have not been cosmetic. Also, as one change overlapped the other, there has been some degree of tension and conflict.

For close to half a century norm-referenced achievement tests were the mainstay of school testing programs. These test batteries, starting with the Stanford in 1923, were developed to provide comparable measures of performance for individual students, evaluated in terms of a single normative group.

Although the dominant mode of testing until the mid-sixties was norm-referenced, another mode had been developing over the years, that of objective-referenced testing. I chose the term objective-referenced purposefully, because much of what is called criterion-referenced or domain-referenced testing is actually objective-referenced, in that the purpose is to assess attainment of objectives. As early as the 1930s,

Ralph Tyler maintained that educational goals should be transformed into measurable objectives and that at the conclusion of an instructional program, pupils should be tested on these objectives. However, it was not until 1969, as Berk (1980, p. 3) notes, that the measurement community was aroused by Popham and Husek's seminal article.

It is my view that the underlying determinant for this interest was the congressional mandate for evaluation of federally funded programs. The 1965 mandate instantaneously created the field of program evaluation. Norm-referenced tests, appropriate for the selection of students, either for remediation or enrichment programs, became instruments for evaluating the programs themselves. Williams and Bank (1980) argue:

> Instead of tests being used primarily to assess individual students, they were now to be used to assess groups of students. Instead of test scores being used, for example, to assign individuals to special education programs, scores were to be used to certify, modify, or eliminate the programs themselves. In short, large-scale testing of students had become a tool for program evaluation (p.3).

The objective-referenced testing movement, I suggest, emerged as program personnel criticized norm-referenced tests for not being closely tied to local instructional programs, thereby being inappropriate for program evaluation.

Swezey, et al., (1972, p. 72) suggest that the primary question of interest in the evaluation of programs is: "Has my instructional program taught what it is suppose to teach?" These authors argue that norm-referenced tests are not appropriate for program evaluation, since overall score change is not necessarily germane to the question of interest. They further argue that objective-referenced tests provide data having relevance to the question.

The educational community has responded enthusiastically to the use of objective-referenced tests, because testing focuses on what a student can do, given particular objectives to accomplish, irrespective of the performance of a nationally representative sample. Berk (1980, p.4) correctly observes, "The explosion of interest that followed in the decade of the 1970s has not abated." Building on Toffler (1981), this objective-referenced testing can be seen to have been both colliding and overlapping with norm-referenced testing. The debate between proponents of one type of measurement or the other continues today, while at the same time both types of measurement have become integral parts of achievement testing programs in schools.

In 1977, another significant measurement innovation became clearly defined. The entire summer issue of the Journal of Educational Measurement was devoted to applications of latent trait models, described by Gary Marco (1977) as a quiet revolution in the field of educational measurement. It has become apparent that controversy exists in this area, but what may not be clear as of yet is that latent trait theory plays an overlapping role because the methodology can be used to join norm-referenced and objective-referenced testing in a customized testing program. Customization of testing programs may be a third wave.

Reasons for Customization

There are two major problems inherent in the use of standardized norm-referenced tests for curriculum evaluation. The first is related to what a test level contains. Given the purpose of spreading student scores, it is obvious that any level of a battery will contain items which measure objectives which are below-grade-level and above-grade-level for the grade in

which the level is administered. Our experience has shown that, although the majority of the items on a test level measure on-grade-level objectives, a substantial number measure below-grade-level and above-grade-level objectives. There are also those items which do not match any local objective and those which use a format which is considered inappropriate for the local population. This comes as a surprise to many people. They may understand the concept of norm-referenced test development, yet they never looked at the implications for curriculum evaluation. If this misunderstanding did not exist, how could anyone talk in terms of measuring the degree to which a grade-level program meets its objectives by administering a norm-referenced test? The second problem is related to what the test does not contain. Our experience has shown that the percentage of local grade-level objectives which are not tested at all on a test level may be as high as 40 percent. Given the fact that a test level may contain an inadequate number of items to assess attainment of a local objective and the fact that many objectives are not assessed by the test, it is obvious that a test level can be inadequate for evaluating a grade-level program.

Solutions Provided by Customization

The replacement of norm-referenced tests with objective-referenced tests has been proposed as a solution to the problems inherent in using norm-referenced tests for program evaluation. Certainly, objective-referenced tests designed to measure specified program objectives can be a solution. However, this solution creates its own problem. The notion of a local and national comparison is heavily entrenched in both the educational and public consciousness, resulting in an expressed concern

to continue norm-referenced testing. Consequently, a school system would have to mount two different testing programs if program evaluation with objective-referenced tests and normative assessment of students were to occur.

A number of solutions for using norm-referenced testing more effectively for program evaluation now exist. The Instructional Dimensions Study (Cooley & Leinhardt, 1980) is a striking example. The procedures involved, however, are both costly and time-consuming. It is not likely that a local school district would have sufficient manpower or time to use the procedures on a large scale.

Palm Beach County Solution: Customized Testing Program

The solution proposed in this symposium entails the multiple use of standardized test items and the supplementation of selected standardized test items with items drawn from a locally-developed bank. Rasch methodology is key to the customization process to be discussed in this symposium.

In developing a new testing program, we operated under the following assumptions.

1. The administration of a norm-referenced achievement test battery is essential. The school district uses the traditional norm-referenced score as a yardstick to compare general performance of local students to that of the national sample and for selection of students for various programs.

2. The norm-referenced test score is inadequate for assessing student performance on the local curriculum. An additional

kind of norm-referenced interpretation would be used for program evaluation. A curriculum- referenced norm score wold be interpreted as comparison of local student performance with national per- formance on only those items which represent the content of the curriculum to which they have been exposed.

3. The norm-referenced test, in and of itself, is inadequate for program evaluation. Objective-referenced testing is necessary for full evaluation of grade level programs and is most appropriate for elementary reading and mathematics programs which have specified objectives.

The customized testing program was developed to meet these assumptions. The Stanford Achievement Test, 7th edition (SAT/7), is to be administered intact and locally-developed tests administered to supplement certain SAT subtests. Customized scoring procedures enable the district to produce three types of scores in addition to the traditional norm-referenced score. A curriculum norm-referenced score for on-grade-level students is based on those items which measure below- and on-grade-level objectives, eliminating from calculation nonmatching and above-grade-level items. Objective-referenced scores are based on both the on-grade-level SAT items and the supplementary items. A total score is also produced for this combined set of items.

The remainder of this symposium will develop the concepts and procedures related to customization.

Lan/209(1a)

REFERENCES

Berk, R. A. "Introduction" in Berk, R.A. (Ed.) Criterion-referenced measurement: The state of the art. Baltimore, Md.: The John Hopkins University Press, 1980.

Cooley, W. W. and Leinhardt, G. The Instructional Dimensions Study, Educational Evaluation and Policy Analysis, 1980, 2 (1), 7-25.

Marco, G. L. Item characteristic curve solutions to three intractable testing problems. Journal of Educational Measurement, 1977, 14, 139-160.

Popham, W. J. & Husek, T. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Swezey, R. W., Pearlstein, R.B., Ton, W. H., and Marabella, A. Contemporary views of criterion-referenced testing. U.S. Army Research Institute for the Behavioral and Social Sciences, 1979.

Toffler, A. The third wave. New York: William Morrow & Co., Inc., 1981.

Williams, R. C. & Bank, A. A school district linking of evaluation and testing with instruction: A problem of attention. Paper presented at the meeting of the Evaluation Research Society, Arlington, VA., November 21, 1980.

USING THE RASCH MODEL TO DEVELOP CURRICULUM-REFERENCED NORMS

R. ROBERT RENTZ

PSYCHOLOGICAL CORPORATION

Presented at the

Annual Meeting of the

National Council on Measurement in Education

New York City

March, 1982

Using the Rasch Model to Develop Curriculum-Referenced Norms

Schools choose a nationally standardized achievement test for many reasons, but certainly one reason is a desire to compare the performance of their students with a nationally representative sample of students on content that represents those broad areas generally taught at particular grades to most students. School systems have long been advised by testing experts to use tests that best match the objectives of their curriculum. Yet, as school systems insist on higher degrees of comparability, it becomes more difficult to find acceptable standardized, norm-referenced achievement batteries that satisfy these standards. A test series may fail to be adopted when only a few items do not match curriculum objectives. A dilemma arises when the desire for norms is frustrated by a less than desirable content/curriculum match. This paper develops the concept of Curriculum-Referenced Norms as a solution to this dilemma.

I will first assert a proposition I believe to be self-evident and then continue as if it were true: fourth graders across our country are generally taught about the same content from one school system to another. The same is true for first graders, second graders, etc. In other words, there is much uniformity in the content of schooling. The presence of such uniformity justifies the promulgation of nationally standardized achievement tests. The authors and publishers of these tests go to great lengths to define their tests' content so that each has the most in common with the largest number of schools. Any other behavior on their part would be contrary to their self-interests. While this emphasis on matching test content to curriculum con-

tent might work well on the whole, in the particular, a given school system will sometimes find a less than perfect match between a test's content and the content of its own curriculum. For example, there may be significant segments of a particular school system's curriculum which the test covers minimally or not at all. Moreover, the test might include items which are not part of that school system's curriculum either at the particular grade being tested or at any grade.

To the extent such mismatches occur, they might be expected to occur proportionately in the standardization sample, and thus the norms would reflect varying degrees of match between test content and curriculum, assuming, of course, a representative sample for standardization. The two sources of mismatch, call them content-not-taught and content-not-tested, probably occur with a different frequency. Content-not-tested is likely to be a more frequent source of mismatch than content-not-taught simply because of the previously mentioned effort to focus test development on curriculum uniformity. On the other hand, content-not-tested is not likely to have a substantial effect on test scores unless the content-not-tested in some way depresses performance relative to the content the test is, in fact, measuring. For example, suppose we have three schools; one has a curriculum consisting of Reading, Math and Science, another Reading, Math and Social Studies, while the third teaches Reading, Math and Sex Education. If, for example, a test is then given in Reading and Math, the presence of Science, Social Studies and Sex Education in the schools' curricula do not affect the interpretation of scores in Reading and Math unless, of course, time is taken away from instruc-

tion in Reading and Math to study Sex. In that case, scores for school 3 might be depressed and securing that information is precisely one of the reasons for testing. A single test battery cannot cover all that schools do; but that does not mean useful, valid information cannot come from tests which cover only a few important areas.

On the other hand, content included in the test that pupils have not had an opportunity to learn (i.e., content-not-taught) can have a serious impact on test scores. Such a statement seems obvious, but its implication with respect to the comparison of an examinee's performance with that of the norms group's may not be so obvious.

First, if it were reasonable to assume that content-not-taught was randomly present to varying degrees for all students in the norms group, then one might argue that this source of score differences does not introduce a score bias in either a positive or negative direction but merely adds to a general error of measurement component, reducing reliability. Consider the hypothetical example of a 100-item test and an examinee norms group in which each examinee has been taught a different combination of from 95 to 100 items such that the mean number of items-not-taught is 2 and the distribution of items-not-taught is normal. Clearly, the effective test length has been reduced by an average of 2 items, more for some examinees, less for others. That is, examinees of maximum ability cannot obtain the identical maximum raw score. As long as an examinee to be compared with this norm's group is from the same population with an expected value of 2 items-not-taught, no bias is introduced.

A problem arises in the above case when, for an examinee or a

school system, there is a significantly greater amount of content-not-taught than the expected amount in the norms group. The scores of such a school system would be depressed relative to the norms even if that school system performed identically to the norms group on every item except those items from the content-not-taught set. Furthermore, there are two different interpretations of such results. If the content-not-taught arises through oversight, neglect, poor instruction, poor learning, etc., then the test has done its proper job in revealing such deficiencies; the scores are not biased and the system can set about correcting the deficiencies. On the other hand, if the system has deliberately chosen to exclude certain content for many justifiable reasons, then the scores would still be depressed and the interpretation that performance was "lower than it should be" would be misleading. The school system might legitimately point out that a fairer comparison should include only that content for which they had assumed responsibility.

I should emphasize here that I am speaking of two different norm-referenced interpretations, both of which are useful but contain different information. In one case the comparison is with a national sample on content either assumed to be common or expected to be common. In the other case, the comparison is with a national sample on content known to be common. I call the norms in the latter case Curriculum-Referenced Norms (CRNs) and will now turn my attention to ways they can be obtained.

Calculating Curriculum-Referenced Norms

For the discussion that follows, assume a school system plans to

administer a nationally standardized, norm-referenced achievement test containing a mathematics subtest of 100 items which includes 5 items on an objective that has not been taught. How could the system obtain normative information on the 95-item test that represents content common to the system's students as well as the norms sample?

One way to obtain some information is to use national item p-values (i.e., the proportion of the norms group correctly answering each item). The procedure involves summing the 95 national p-values, yielding the mean raw score of the norms group on the 95-item test. The school system could then compare its students' performances against this national mean. Tests would have to be rescored to account for the five fewer items and the only normative information available would be whether the examinee was above or below the national average. This information might suffice for some purposes.

More information could be obtained if, in addition to p-values, the publisher provides item discrimination indexes, in the form of point-biserial correlations. An approximate standard deviation can be calculated by summing, over items, the product of the discrimination index and the item standard deviation. Armed with the national mean and national standard deviation of the 95-item test and, further, assuming a normal distribution of scores, estimated percentile ranks can be calculated based on a cumulative normal distribution, and tables of raw scores to percentile ranks could be produced. For a variety of reasons, such an approximation would probably not be very accurate, particularly for extreme percentile ranks.

A third procedure, and probably the most accurate, would require

the publisher to recalculate all norms tables based on the empirical norms data for the 95-item test. That might be relatively simple for one test, but if there were multiple subtests and multiple levels and if the school system wanted percentile ranks, stanines and grade equivalents, the scope and cost of the task would be very difficult to justify.

In summary, these three procedures either provide too little information, are too inaccurate or are too expensive. The difficulty seems to lie in the way norms tables have to be developed for a multilevel test battery. These procedures are quite complex, time consuming and expensive, even after the data themselves have been collected.

In addition to the procedures mentioned above, there are at least two solutions to the calculation problem in the context of latent trait, or item response, theory. One solution is the use of the compound binomial error models discussed in Chapter 23 of Lord and Novick (1968). These models use item response curve parameters to obtain estimated observed-score distributions for a test without administering that test to any norms group. The procedure is computationally difficult but can be programmed; however, it only yields a score distribution from which norms tables would still have to be obtained. In this sense it differs little from the recalculation method using empirical data.

If we change the problem from a norms development problem to an equating problem, a relatively simple solution emerges. If we could equate the 95-item test to the 100-item test, then norms could be found directly from the tables for the 100-item test without further

calculation. A further simplification arises if the latent trait model's ability scale has been used as the basis for norms development, thus avoiding a raw score-to-raw score equating transformation which is a major source of equating error [see Rentz & Bashaw's (1977) discussion of "assignment error"]. A still further simplification is obtained when the Rasch Model, in contrast to more complex models, has been used for test development. The Rasch Model requires only the calculation of a new raw score to ability scale conversion table, whereas more complex models would require the reestimation of ability for each examinee. The specific details of how the CRNs are derived are described by Gramenz in a later symposium presentation.

That is the easy part; now for the hard part. Let's look carefully at what we have done. First of all, on a theoretical level, if a test has been developed using Rasch Model procedures and a bank of calibrated items are available, any subset of these calibrated items can be used to compose a test with the result that scores expressed on the ability scale are automatically equated to scores from any other subset of items. This is true for all latent trait models and is the equating principle inherent in these models. This theoretical principle means that in practice we would expect our 100-item test to yield the same ability scale values for examinees as our 95-item test. Thus, if the ability values are equal and norms are based on the ability scale, then the norms are equal, and there would be no difference between the full-length test norms and the Curriculum-Referenced Norms. How is it possible, then, to ever get different results for CRNs?

The answer depends on how the content-not-taught is affected by the instructional process. Throughout this paper I have assumed that it is fairly straightforward to identify items in a test that reflect content-not-taught. Actually, this identification is a bit tricky. For there to be a difference in the ability values for these two tests, the relative difficulty of items deleted as content-not-taught must be different in the local group being tested compared with the norms group. For example, Figure 1 illustrates the theoretically expected relationship between item difficulty values obtained from two different calibration situations. The two calibration situations could involve two different groups of examinees or the same group at two different times. This plot represents item difficulty values for the national sample compared with a local sample. Figure 1 shows that the relative item difficulties are expected to be constant. Figure 2, in contrast, shows what is likely to happen to the relative item difficulties when there is a subset of items representing content-not-taught. These items have shifted in difficulty compared to the total set of items. These five items-not-taught are more difficult for the local group, as you might expect. Furthermore, the amount of shift is greatest for items that were the easiest for the national sample. While it is possible for the set of items-not-taught to be items difficult for the national group, such a situation would be practically indistinguishable from that shown in Figure 1. The largest effect occurs when the items are easy for the national group. This latter case seems more likely and is assumed for the remainder of this discussion. In other words, the local sample tends to correctly answer

the easier items less frequently than expected based on the national sample's performance. Removing the five items has the effect of making the resulting shorter test more difficult for the national group; but since the local examinees are answering correctly a proportionately greater number of items from a more difficult test, the estimate of their ability has to be greater than the estimate obtained for the full-length test.* Thus, for the ability values on the two tests to be different, some pattern analogous to that shown in Figure 2 must be present. If the school system identifies items as content-not-taught and those items behave like those in Figure 1, then there will be no difference between CRNs and regular norms.

I'm not sure I know what all this means, but three possible explanations do occur to me. One, increasing proficiency with respect to parts of a domain increases proficiency for the entire domain. For example, an increase in performance on addition problems might lead to improvement in subtraction, even in the absence of subtraction instruction. Two, the instructional process itself is singularly ineffective; there is no difference in the impact on content-not-taught and content-taught. Finally, the school system identified the wrong items as content-not-taught. Actually, there is one more explanation, but it is too far-fetched to take seriously; that is that the test has zero validity.

---

*This is one of those intuitively obvious statements, an explanation of which requires another five pages.

Summary

In this paper I have attempted to develop the concept of Curriculum-Referenced Norms as a solution to the problem of a less-than-perfect match between the content of a nationally standardized achievement test and the content of a local school system's curriculum. I have suggested that the computation of such norms is facilitated if the Rasch Model has been the basis for test construction, even though CRNs can be obtained, with considerable effort, using other procedures. Finally, I have tried to sprinkle throughout the paper hints on how such norms can be interpreted. I will eagerly await the outcome of the work in Palm Beach County to see if these ideas, many of which are not particularly new, have any practical merit.

REFERENCES

Rentz, R. R. & Bashaw, W. L. The National Reference Scale for Reading: An Application of the Rasch Model. Journal of Education Measurement, 1977, 14, 161-180.

Lord, R. M. & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison Wesley, 1968, 23.

4:205:8

D/NCME/A

Figure 1:
Expected Relationship Between Difficulty Parameters for Same Items Obtained From Two Calibration Groups

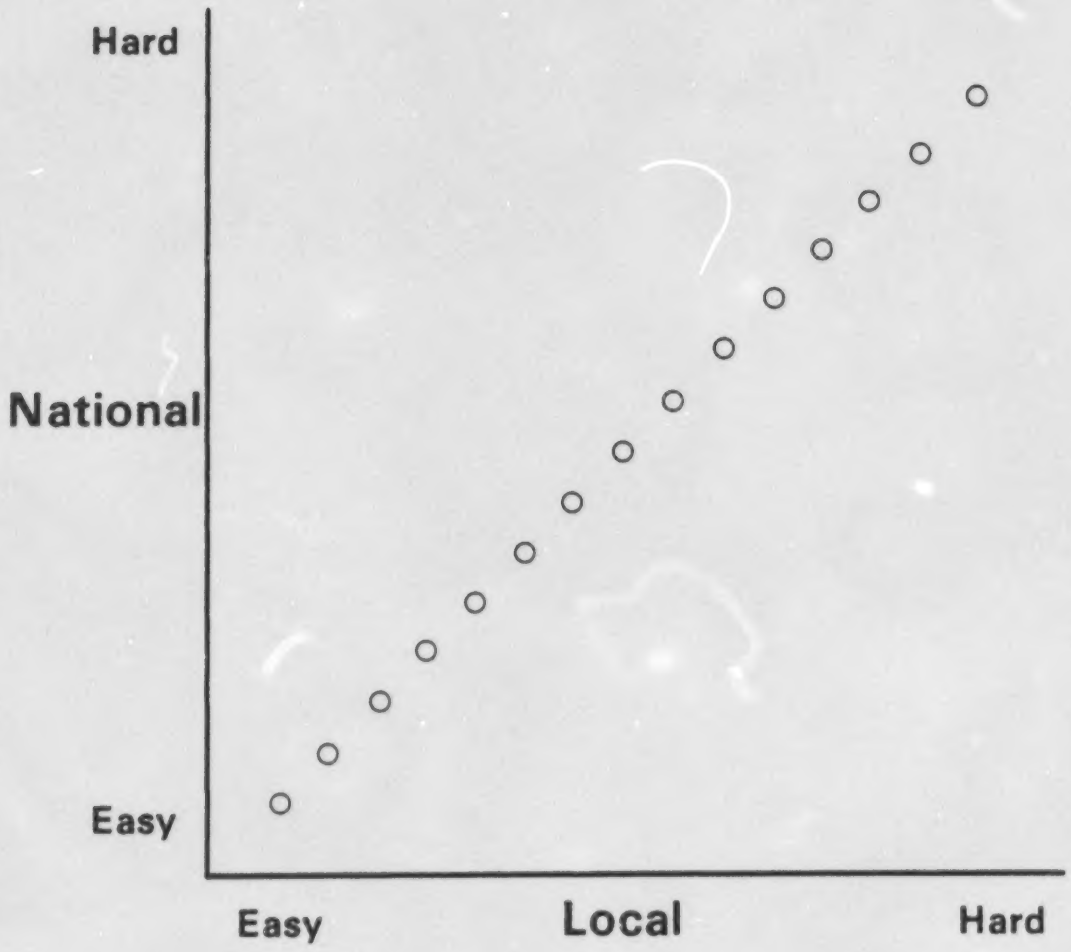
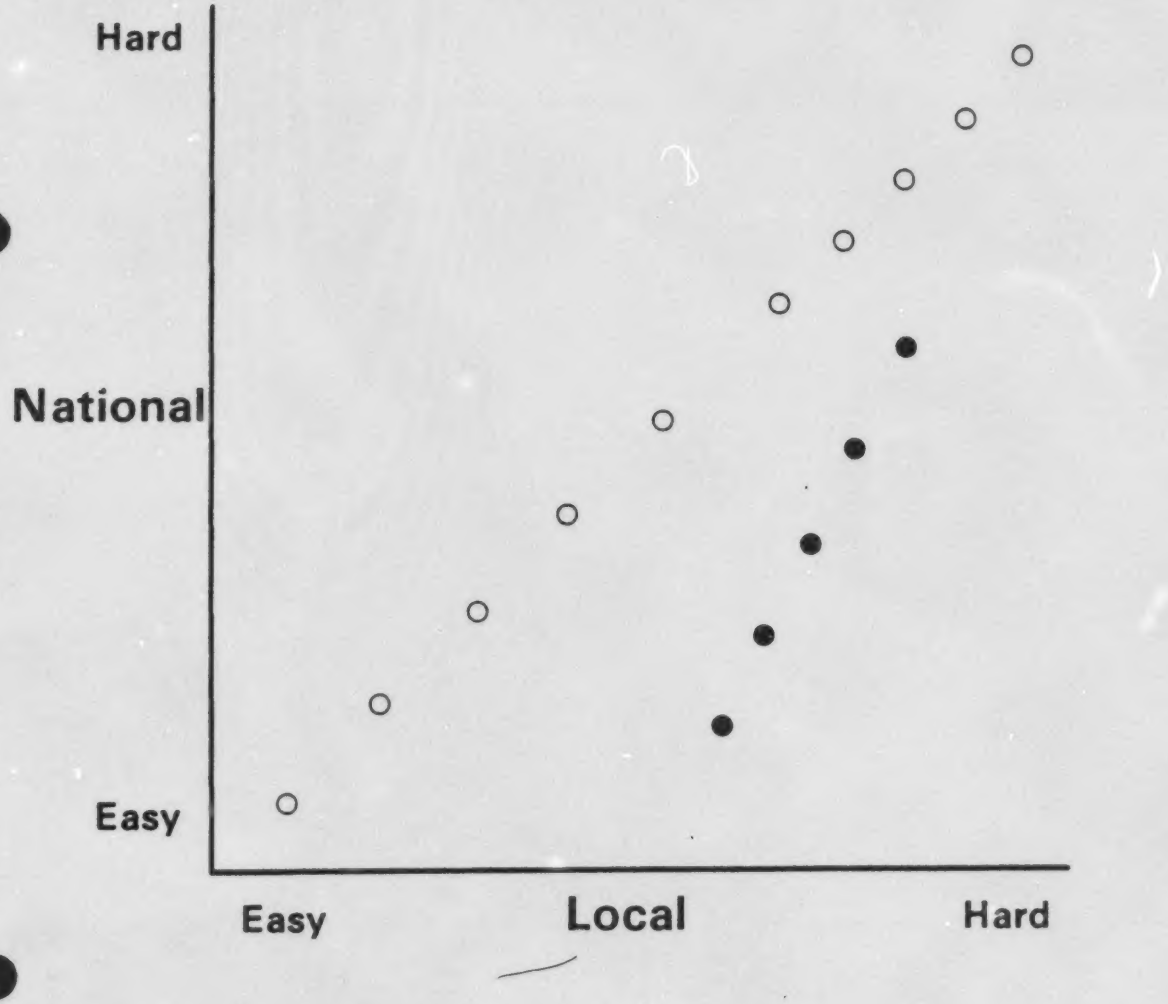


Figure 2:

Relationship Between Difficulty Parameters for Same Items Obtained From Two Calibration Groups When a Subset of Items Is Not Taught



MATCHING STANDARDIZED ACHIEVEMENT TEST ITEMS TO LOCAL CURRICULUM OBJECTIVES

RUTH BOWER

PALM BEACH COUNTY SCHOOLS

Presented at the

Annual Meeting of the

National Council on Measurement in Education

New York City

March, 1982

Matching Standardized Achievement Test Items To Local Curriculum Objectives

Introduction

Mismatch between standardized achievement test items and local curriculum objectives is not a new issue. Considerable literature has been generated over the past several years regarding the lack of fit between curriculum objectives and the content of norm-referenced tests. Tallmadge (1977) notes that, although standardized achievement tests are designed to be curriculum free, this is not the case. Similarly, Green (1979) maintains that one can almost never find a standardized achievement test whose content closely matches local instructional objectives.

Numerous studies of the degree of match between curriculum and test content have been conducted. (Leinhardt and Sewald, 1980, cite the major studies.) Cox and Sterrett (1970) propose that any model for studying the degree of match must include a precise description of curriculum objectives and the coding of each standardized achievement test item with reference to the curriculum. As Koslin, et. al (1979) indicate, this can be a major problem, even when clearly specified local objectives are available, since test publishers do not provide operational definitions of just what skill is being measured by an item. Nonetheless, Leinhardt and Sewald (1970) suggest that overlap information should be incorporated into the interpretation of evaluation results.

Background of the Study

In March, 1981, the Palm Beach County Schools adopted the Stanford Achievement Test, 7th Edition (SAT/7), for district-wide administration. Experience with the 6th Edition, which had been administered since 1973, had convinced us that it was not likely that the content of the 7th Edition would closely match our curriculum in each subject area at every grade level. In fact, we were convinced that no standardized achievement test battery could meet such a criterion. Nevertheless, we needed to know the degree of match.

In the summer of 1981, a team was formed to study the SAT/7 reading and mathematics items which would be administered to students in grades two through six. The 18-member team was composed of math and reading consultants, a testing consultant, elementary principals and assistant principals and elementary teachers. Teachers were selected to provide a balanced representation across the elementary grade levels. SAT/7 reading items were to be matched with objectives in the Ginn reading series and SAT/7 mathematics items with local mathematics objectives, which included objectives taken from the Heath mathematics series.

Procedures

The team was divided into four grade-level groups, working with designated levels of the SAT/7. Team members were provided local curriculum materials and copies of the appropriate levels of the SAT/7. A representative from Psychological Corporation worked with the team, answering questions concerning test construction.

The team spent a considerable amount of time defining the criteria

to be used in the matching process. Clearly, it was imperative that the definitions of such terms as match and mismatch be understood and accepted by all members of the team before assigning grade-level designations to test items.

Each group classified the items on a given test level as either matching an on-grade-level, below-grade-level or above-grade-level objective, or not matching any objective. Then each group presented its classifications to the entire team, and each item was examined again. If, at this point, there still were questions regarding the match or mismatch designations of items, the items were referred back to the group for reconsideration and again presented to the team. After concensus was reached regarding all item designations, a report was compiled delineating the SAT/7 items which matched the local curriculum objectives.

Reading

The team examined each reading item for its match with the Ginn reading objectives. Two factors were considered in determining whether a given item matched an on-grade-level, below-grade-level or above-grade-level objective: the Ginn textbook level in which the objective measured by the item was introduced and the readability level of the item. The readability level was obtained by comparing the correct response for the item (both the stem and the correct response in the case of Reading Comprehension items) to the Ginn Cumulative Word List and the Ginn Scope and Sequence Skills Chart. In those instances where the textbook level and readability level of an item differed, the item was assigned the higher of the two levels. For example, the Ginn textbooks

used in second grade are levels six and seven. If the objective measured by an item was introduced in level seven, but the readability level of the item was level eight, the item was classified as measuring an above-grade-level objective.

Reading items which did not match any objective included in the Ginn reading series were classified as mismatches. An extreme example was the mismatch of elementary curriculum objectives and the Word Study Skills subtest of Intermediate I, Form E. Syllabication is not taught in the Ginn core program in the manner in which it is tested on the Word Study Skills subtest. The approach to syllabication used in the Ginn series should cause students to select incorrect responses to eight of the 24 Word Study Skills items. The remaining 16 items, although measuring syllabication, do not conflict directly with Ginn instruction. Only those items which have double consonants, unstressed vowels standing alone, or any "y" vowell in a syllable standing alone, should elicit an incorrect response from students exposed to the Ginn core program.

To check this assumption, the Intermediate I, Form E, Word Study Skills subtest was administered to 93 students who had just completed fourth grade. The results of the study are shown in Table 1.

Table 1

Results of Exploratory Study of Intermediate I, Form E, Word Study Skills Items

Percentage of Correct Responses	Items Having Correct Response Which Does Not Conflict With Ginn Instruction	Items Having Correct Response Which Conflicts with Ginn Instruction
90-100	1	0
80-89	6	0
70-79	3	0
60-69	5	4
50-59	1	1
40-49	0	3
below 40	0	0

The data indicate that, in those cases where the item is not in direct conflict with the Ginn method of teaching syllabication, student performance spreads across the range of 50 percent to 100 percent correct. Student performance on items which are in direct conflict with Ginn instruction ranges from 40 percent to 69 percent correct. These results supported the team's opinion that the eight items should be classified as mismatches.

Mathematics

Each SAT/7 mathematics item was examined for its match with local mathematics objectives. The matching process was facilitated by the clarity with which the objectives were stated. Consider, for example, the two pairs of objectives and test items shown below. Test item 1 clearly matches objective 1. Test item 2, however, does not match objective 2, since the item does not require the student to measure the length of the pencil.

Objective 1:	The student will solve one-step word problems using multiplication or division skills previously mastered.
Test Item 1:	Jane runs 2 blocks every day. How many blocks will she run in 5 days?
Objective 2:	The student will measure height, width and length of figures using customary or metric units.
Test Item 2:	What metric unit would be used to measure how long a pencil is?

Results

The results of the analysis of the SAT/7 reading items are shown in

Table 2, which contains the classifications of all reading items which matched curriculum objectives at any elementary grade level.

Table 2

SAT/7 Reading Item Match With Ginn Reading Objectives

Grade Level	SAT/7 Level	No. Items	Percent of Items Measuring Grade-Level Objectives: Below	On	Above	Total*
2	Pri. II(E)	121	9	59	32	100
3	Pri. III(E)	114	21	39	37	97
4	Int. I(E)	120	34	33	23	90
5	Int. II(E)	120	47	23	23	93
6	Int. II(F)	120	53	18	20	91

The data clearly reflect large differences in the proportions of reading items on a particular SAT/7 level which match the reading objectives taught in the grade at which that level is administered. For example, 59 percent of the Primary II, Form E, reading items match second grade reading objectives. However, only 18 percent of the Intermediate II, Form F, reading items match sixth grade reading objectives. Fifty-three percent of this level's reading items match below-grade-level objectives and 20 percent of them match above-grade-level objectives. The total percentage of items which match reading objectives at any grade level is very high, ranging from a low of 90 percent at grade four to a high of 100 percent at grade two. In other words, there are very few SAT/7 reading items which do not match the objectives taught in the elementary reading curriculum.

The results of the analysis of the SAT/7 mathematics items are shown in Table 3, which contains the classifications of all mathematics items which matched curriculum objectives at any elementary grade level.

Table 3

SAT/7 Mathematics Item Match With Local Mathematics Objectives

Grade Level	SAT/7 Level	No. Items	Percent of Items Measuring Grade-Level Objectives: Below	On	Above	Total*
2	Pri. II(E)	108	0	81	16	97
3	Pri. III(E)	114	18	45	22	85
4	Int. I(E)	118	26	32	21	79
5	Int. II(E)	118	15	63	9	87
6	Int. II(F)	118	27	60	0	87

As was the case with the reading items, there is substantial variation in the proportion of mathematics items on a particular SAT/7 level which match the mathematics objectives taught in the grade at which that level is administered. Whereas 81 percent of the Primary II, Form E, mathematics items match second grade mathematics objectives, only 32 percent of the Intermediate I, Form E, mathematics items match fourth grade mathematics objectives. Twenty-six percent of this level's mathematics items match below-grade-level objectives and 21 percent of them match above-grade-level objectives. The total percentage of items which match mathematics objectives at any grade level ranges from a low of 79 percent at grade four to a high of 97 percent at grade two.

Summary

Team members completed their task in twelve days. Each reading and mathematics item on the SAT/7 levels to be administered in grades two through six had been analyzed. The preliminary step in customizing an achievement testing program had been accomplished.

REFERENCES

Cox, R. C. & Sterrett, B. G. A Model for Increasing the Meaning of Standardized Test Scores. Journal of Educational Measurement, 1970, 7, 227-228.

Green, D. R. "Individual Comments," in Testing, Teaching and Learning. National Institute of Education, Washington, D. C., 1979, 29-31.

Koslin, B. L., Koslin, S., & Zeno, S. "Toward an Effectiveness Measure in Reading," in Testing, Teaching and Learning. National Institute of Education, Washington, D. C., 1979, 311-334.

Leinhardt, G. & Sewald, A. Overlap: What's Tested, What's Taught. National Institute of Education, Washington, D. C., 1980.

Tallmadge, G. Kasten. The Joint Dissemination Review Panel IDEABOOK. National Institute of Education, Washington, D. C., 1977.

AN EXPLORATORY STUDY OF THE CONCEPT OF CURRICULUM-REFERENCED NORMS USING THE STANFORD ACHIEVEMENT TEST, SIXTH EDITION

GARY W. GRAMENZ
ROBERT C. JOHNSON
BERNARD G. JONES

PALM BEACH COUNTY SCHOOLS

Presented at the

Annual Meeting of the

National Council on Measurement in Education

New York City

March, 1982

An Exploratory Study of the Concept of Curriculum-Referenced Norms Using the Stanford Achievement Test, Sixth Edition

Introduction

The purpose of this study is to conduct an exploratory investigation of the concept of Curriculum-Referenced Norms (CRNs), using the Primary I/II reading and mathematics items of the Stanford Achievement Test, Sixth Edition (SAT/6). The study was designed to serve as the basis for making tentative generalizations about the potential utility of CRNs and for shaping future research activities involving the reading and mathematics items of the Stanford Achievement Test, Seventh Edition (SAT/7).

In an earlier symposium presentation, Rentz identified two potential sources of mismatch between the content of a nationally standardized achievement test and the content of a given school system's curriculum. One source of mismatch occurs when the test's content does not measure a portion of the school system's curriculum (i.e., content-not-tested). The second source of mismatch occurs when a portion of the test's content measures objectives which are not part of the school system's curriculum (i.e., content-not-taught). As Rentz indicated, the content-not-taught source of mismatch would, under certain conditions, depress a school system's test scores relative to the norms. This could result in the misleading interpretation that students did not perform as well as they should have. The use of Curriculum-Referenced Norms is a

potential means of eliminating the content-not-taught source of mismatch and, hence, such a misleading interpretation of a school system's test scores relative to the norms.

The major assumption underlying the concept of CRNs is that students should perform better on a test which measures only content-taught than on a test which measures both content-taught and content-not-taught. This study is an initial investigation of this assumption.

Procedures

The first step in our exploratory investigation of the major assumption underlying the concept of Curriculum-Referenced Norms was to match the SAT/6 Primary I/II reading and mathematics items with local curriculum objectives. Two different matching processes were used for reading items. The first process (Method A) was virtually identical to that described by Bower in an earlier symposium presentation. The second process (Method B) differed from the first in that both the correct response and the distractors for each item were considered in arriving at an average readability level. The matching process used for mathematics items also was virtually identical to that described by Bower.

The national Rasch calibrations of the SAT/6 reading and mathematics items, based on the 1972 standardization, were obtained from Psychological Corporation. The calibrations were merged with the item classifications to distinguish the content-taught (i.e., curriculum-referenced) subsets of items from the content-not-taught subsets of items. The results of this match were processed through Psychological Corporation's

TRIDELT computer program, which created raw score-to-Rasch ability estimate conversion tables for Primary I/II Total Reading, Total Mathematics and the corresponding curriculum-referenced subsets. The ability estimates for Total Reading and Total Mathematics were then regressed on the appropriate SAT/6 scaled scores. The regression coefficients obtained from this analysis were processed through TRIDELT, which transformed the ability estimates into Rasch scaled scores.

The degree of linearity between the Rasch ability estimates and the SAT/6 scaled scores also was determined from the regression analysis. The correlation coefficients which were obtained (0.98 for Primary I mathematics and Primary II reading; 0.99 for Primary I reading and Primary II mathematics) confirmed the presence of a high degree of linearity between the two sets of scores. Confirmation of such a high degree of linearity enabled us to use the SAT/6 scaled score-to-percentile rank conversion tables in estimating the gains or losses on the curriculum-referenced subsets.

Student item responses from the most recent (Spring, 1980) district administration of the SAT/6 were processed through a local program, which computed raw scores for Total Reading, Total Mathematics and the curriculum-referenced subsets. The TRIDELT conversion tables were used to transform the raw scores to Rasch scaled scores, after which mean Rasch scaled scores were computed for the various item groups. The latter values enabled us to ascertain whether students performed better on the curriculum-referenced subsets than they did on the total item groups.

To determine whether student gains or losses on the curriculum-referenced subsets varied with the SAT/6 Total Reading and Total Mathe-

matics scores, mean Rasch scaled score differences between the subset and total scores were calculated for stanines one through nine.

In addition, to determine whether any subset of items would yield results similar to those of the curriculum-referenced subsets, several random samples, identical in size to those of the subsets, were drawn.

Finally, to help explain why student performance on a given curriculum-referenced subset improved or declined relative to the total item group, the University of Chicago's FORCAL computer program was used to generate local item difficulties based on the Spring, 1980, administration of the SAT/6. These values were plotted against the national item difficulties from the 1972 standardization, and the plots examined for items showing substantially different local and national calibrations.

Results

The mean Rasch scaled scores for the various SAT/6 reading and mathematics item groups are shown in Table 1. The Primary I curriculum-referenced reading means for Method A and Method B were approximately one Rasch scaled score point above those of the corresponding Total Reading means. The differences in the mean scores resulted in student gains of one to two percentile rank points.

The results of the Primary II reading analysis were the opposite of those observed for Primary I. Whereas the differences in the mean scores for Primary I favored the curriculum-referenced subsets, the differences in the mean scores for Primary II favored Total Reading by less than one scaled score point for both Method A and Method B.

The results of the Primary I/II mathematics analyses showed a pattern similar to that of the corresponding reading analyses. First grade students evidenced a slight gain on the curriculum-referenced subset, while second grade students showed a slight loss.

The results of the Primary I and Primary II reading analyses by stanine are shown in Table 2 and Table 3, respectively. The analyses for Method A and Method B reflect a similar pattern of mean scaled score differences: students in the upper and lower stanines evidenced the greatest losses on the subsets, while those in the middle stanines showed either the smallest losses (Primary II) or modest gains (Primary I).

The results of the Primary I/II mathematics analyses by stanine are shown in Table 4. These data reflect a second distinct pattern: students in the lower four or five stanines evidenced losses on the subsets, while students in the upper four or five stanines showed gains.

Hence, the stanine analyses revealed two general patterns: for Primary I/II reading, students in the middle stanines fared better on the curriculum-referenced subsets than did students in the upper and lower stanines while, for Primary I/II mathematics, students in the upper four or five stanines fared better on the subsets than did those in the lower stanines.

The mean Rasch scaled score differences for Total Reading, Total Mathematics and the appropriate random samples of items are shown in Table 5. The results of the data analyses indicated that randomly selected samples of items yielded mean scores approximately equivalent to those obtained for the total item groups. The magnitude of the

mean differences ranged from a low of 0.01 scaled score points (Primary II reading, Sample 2) to a high of -0.51 scaled score points (Primary I reading, Sample 2). Furthermore, the mean scaled score differences between the random samples and the total item groups generally were smaller than those between the curriculum-referenced subsets and the total item groups (see Table 1).

The bivariate plots of the national vs. local SAT/6 item difficulties are shown in Figures 1-4. The plot of the Primary I reading item difficulties (Figure 1) revealed 17 outliers (i.e., items whose national and local difficulties differed by 0.5 logits or more). Nine items fell below the 0.5 logit cutoff, indicating they were more difficult for local students than they were for the national sample. Four of these items (all from the Word Study Skills subtest) were classified as content-taught by Method A and none by Method B. Eight items were located above the cutoff, indicating they were less difficult for local students than they were for the national sample. Three of these items were classified as content-taught by Method A and one by Method B. Four of the eight items were from Reading Part A, one from Reading Part B and three from the Word Study Skills subtest.

The plot of the Primary II reading item difficulties (Figure 2) showed 13 outliers below the lower boundary, all but one of which, including a string of ten consecutive items, measured syllabication. Ten of these items were classified as content-taught by Method A and six by Method B. In addition, 14 outliers were located above the upper boundary, six of which were classified as content-taught by Method A and four by Method B. Eight of these items were located in Reading Part A, four in Reading Part B and two in the Word Study Skills subtest.

The plot of the Primary I mathematics item difficulties (Figure 3) revealed seven outliers below the cutoff, four of which were classified as content-taught. Five more outliers, four of which were classified as content-taught, were located above the cutoff. The 12 outliers were about evenly divided between the two mathematics subtests.

The plot of the Primary II mathematics item difficulties (Figure 4) showed five outliers below the lower boundary, four of which were classified as content-taught. Two more outliers, both classified as content-taught, were located above the upper boundary. The seven outliers were about evenly distributed across the three mathematics subtests.

The differences between the local and national mean item difficulties for the curriculum-referenced subsets and the randomly selected samples are shown in Table 6 and Table 7, respectively. These differences account for student gains or losses on the various item subsets. For example, the 1.20 scaled score point gain on the Primary I reading, Method B, subset of items (Table 6) was due to the fact that the mean difficulty of these items was 0.09 logits more difficult for the national sample than it was for local students. Conversely, the 0.51 scaled score point loss on the second random sample of Primary I reading items (Table 7) was due to the fact that the mean difficulty of these items was 0.06 logits less difficult for the national sample than it was for local students.

Summary and Conclusions

The overall results of the study revealed generally small mean Rasch scaled score differences between the curriculum-referenced subsets

and the total item groups. The differences in mean scores were due to the differences between the national and local difficulties of the items comprising the subsets. The results of the reading analyses by stanine showed that students in both the upper and lower stanines incurred relatively large losses on the curriculum-referenced subsets. Further research should be conducted to determine whether the use of CRNs will adversely affect students within particular score ranges.

Because of the limited scope and inconclusive results of the preliminary research reported here, we do not feel that any meaningful conclusions regarding the practical utility of the concept of Curriculum-Referenced Norms can be drawn. A more comprehensive investigation of the major assumption underlying the concept of CRNs will be conducted this summer. The study will involve four levels of the SAT/7 reading and mathematics items (Primary II through Intermediate II). In addition, a second student classification scheme, based on progress in the language arts curriculum and the mathematics curriculum, will be utilized. Students in grades two through six will be identified as performing below-grade-level, on-grade-level or above-grade-level just prior to the administration of the SAT/7. Hopefully, use of this classification scheme, in conjunction with a replication of the initial study using four levels of the SAT/7, will provide a definitive answer to the question of whether students do, indeed, perform better on a test which measures only content-taught than on a test which measures both content-taught and content-not-taught.

Table 1

Mean Rasch Scaled Scores for Total Reading, Total Mathematics and Curriculum-Referenced Subsets:
SAT/6 Primary I and Primary II

SAT/6 Content Area	No. Cases	Subset No. Items	Subset SS Mean	Subset SD	Total No. Items	Total SS Mean	Total SD	SS Diff.*	SD	Percentile Rank Total	Percentile Rank Subset	Est. %ile Diff.
Pri. I Reading+												
Method A	4,621	49	110.20	15.50	147	109.17	13.81	1.02	6.25	35	37	+2
Method B	4,409	26	109.23	14.75	147	108.03	12.86	1.20	6.89	34	35	+1
Pri. II Reading+												
Method A	5,011	114	127.97	18.90	158	128.57	18.71	-0.59	2.88	44	44	0
Method B	5,000	75	127.63	18.37	158	128.46	18.58	-0.82	4.79	44	43	-1
Pri. I Math	4,886	41	122.89	9.10	64	122.63	8.18	0.25	2.60	66	68	+2
Pri. II Math	5,007	76	133.87	10.49	100	133.91	10.03	-0.04	2.05	60	60	0

+The discrepancies in the number of cases and in the mean scaled scores for the total item set between Method A and Method B are an artifact of perfect scores being eliminated in the Rasch calibrations.

*Subset SS - Total SS

Table 2

Mean Rasch Scaled Score Differences for Total Reading and Curriculum-Referenced Subsets, by Stanine: SAT/6 Primary I

	Primary I Reading					
	Method A			Method B		
Stanine	N	Mean SS Diff.*	SD	N	Mean SS Diff.*	SD
1	78	-5.45	4.78	78	-4.77	6.18
2	431	-3.87	4.34	431	-2.87	5.56
3	737	-1.23	3.71	737	-0.66	5.08
4	995	1.87	4.25	995	1.60	5.13
5	1,061	3.80	5.67	1,051	4.38	6.60
6	722	4.37	6.78	669	4.44	6.93
7	358	1.08	6.45	294	-0.46	6.25
8	153	-4.43	5.52	103	-7.98	5.13
9	86	-12.09	4.68	51	-16.08	4.46
Total	4,621	1.02	6.24	4,409	1.20	6.89

*Subset SS - Total SS

Table 3

Mean Rasch Scaled Score Differences for Total Reading and Curriculum-Referenced Subsets, by Stanine: SAT/6 Primary II

	Primary II Reading					
	Method A			Method B		
Stanine	N	Mean SS Diff.*	SD	N	Mean SS Diff.*	SD
1	254	-1.41	1.84	254	-0.75	3.51
2	483	-1.37	1.77	483	-0.21	2.87
3	615	-0.68	1.81	615	0.13	3.13
4	908	-0.23	1.90	908	0.21	3.87
5	1,060	-0.06	2.50	1,060	-0.92	4.81
6	749	-0.43	3.30	749	-1.75	5.61
7	430	-0.44	3.80	430	-1.08	6.21
8	245	-0.80	4.61	246	-1.85	5.96
9	267	-2.07	4.77	255	-3.43	6.43
Total	5,011	-0.59	2.87	5,000	-0.82	4.79

*Subset SS - Total SS

Table 4

Mean Rasch Scaled Score Differences for Total Mathematics and Curriculum Referenced Subsets, by Stanine: SAT/6 Primary I and Primary II

	Mathematics					
	Primary I			Primary II		
Stanine	N	Mean SS Diff.*	SD	N	Mean SS Diff.*	SD
1	97	-1.34	1.80	47	-1.11	1.26
2	231	-0.89	2.21	200	-0.80	1.43
3	431	-0.29	2.07	476	-0.62	1.31
4	823	-0.22	2.03	886	-0.19	1.37
5	1,009	0.06	2.18	1,003	-0.15	1.68
6	853	0.55	2.39	943	0.19	1.84
7	624	0.60	2.65	555	0.15	2.29
8	491	1.15	3.39	515	0.62	2.60
9	327	1.28	3.87	382	0.09	3.64
Total	4,886	0.25	2.60	5,007	-0.04	2.05

*Subset SS - Total SS

Table 5

Mean Rasch Scaled Score Differences for Total Reading, Total Mathematics and Random Samples of Items: SAT/6 Primary I and Primary II

SAT/6 Content Area	No. Cases	Subset No. Items	Subset SS Mean	Subset SD	Total No. Items	Total SS Mean	Total SD	SS Diff.*	SD
Pri. I Rdg. - Method A									
Sample 1	4,783	49	109.88	14.09	147	110.08	14.48	-0.20	3.66
Sample 2	4,780	49	109.58	14.95	147	110.09	14.53	-0.51	4.01
Pri. II Rdg. - Method A									
Sample 1	5,014	114	128.61	18.55	158	128.58	18.71	0.03	1.99
Sample 2	5,008	114	128.54	18.55	158	128.53	18.65	0.01	2.12
Pri. I Math.									
Sample 1	4,895	41	122.79	8.60	64	122.68	8.23	0.11	1.83
Sample 2	4,891	41	122.71	8.50	64	122.66	8.22	0.05	1.95
Pri. II Math.									
Sample 1	5,014	76	134.13	10.36	100	133.96	10.11	0.18	1.52
Sample 2	5,014	76	134.02	9.99	100	133.96	10.10	0.07	1.52

*Subset SS - Total SS

Table 6

Mean Differences Between National and Local Reading and Mathematics Item Difficulties for Curriculum-Referenced Subsets: SAT/6 Primary I and Primary II

SAT/6 Content Area	Mean d_i Difference (National - Local)	Mean SS Difference
Pri. I Reading		
Method A	0.04	1.02
Method B	0.09	1.20
Pri. II Reading		
Method A	-0.05	-0.59
Method B	-0.06	-0.82
Pri. I Math	0.03	0.25
Pri. II Math	0.01	-0.04

Table 7

Mean Differences Between National and Local Reading and Mathematics Item Difficulties for Random Samples: SAT/6 Primary I and Primary II

SAT/6 Content Area	Mean d_i Difference (National - Local)	Mean SS Difference
Pri. I Rdg. - Method A		
Sample 1	0.02	-0.20
Sample 2	-0.06	-0.51
Pri. II Rdg. - Method A		
Sample 1	0.00	0.03
Sample 2	0.01	0.01
Pri. I Math		
Sample 1	-0.04	0.11
Sample 2	-0.01	0.05
Pri. II Math.		
Sample 1	0.02	0.18
Sample 2	0.01	0.07

FIGURE 1

NATIONAL VS. LOCAL RASCH ITEM DIFFICULTIES FOR SAT/6 PRIMARY 1: TOTAL READING

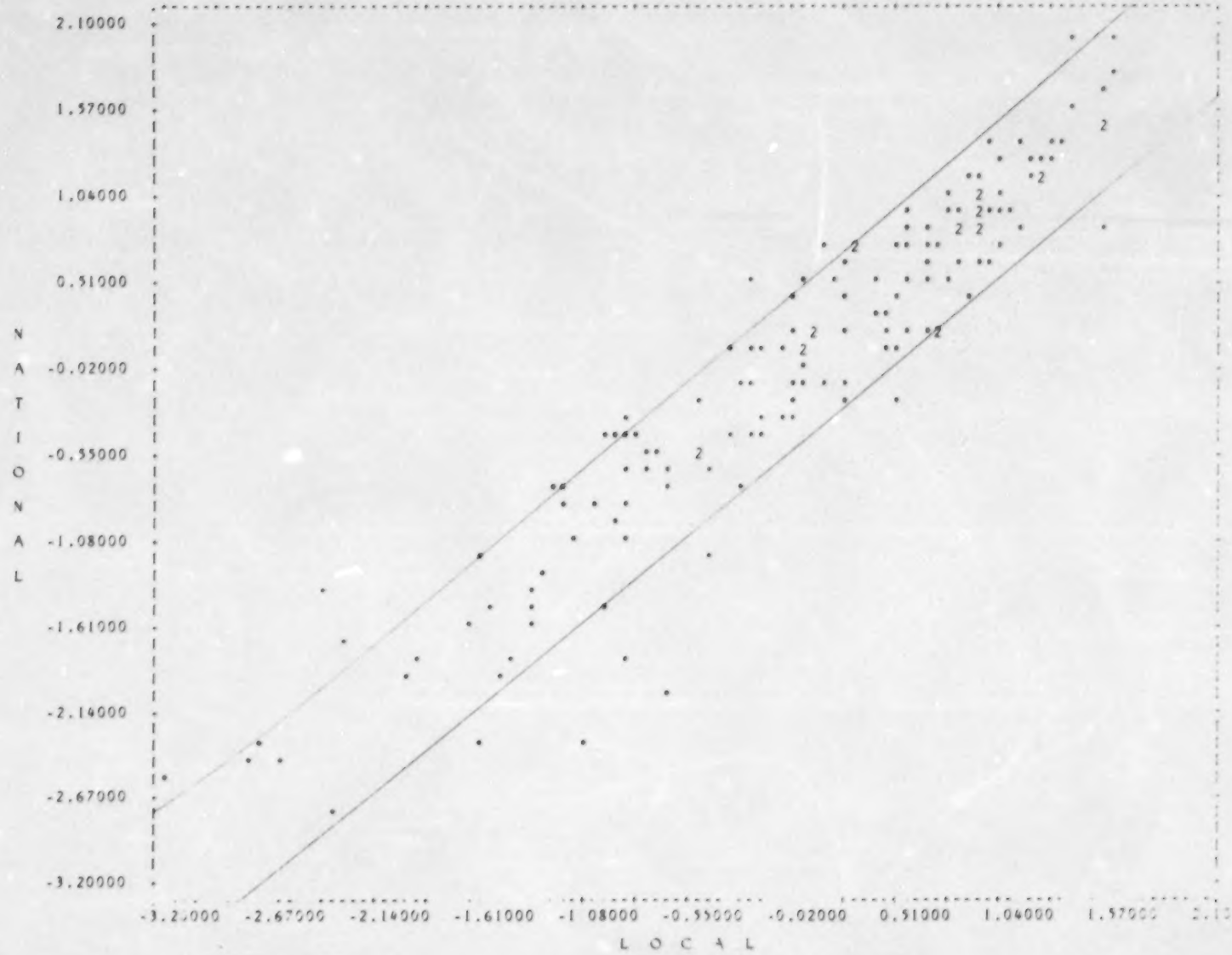


FIGURE 2

NATIONAL VS. LOCAL RASCH ITEM DIFFICULTIES FOR SAT/6 PRIMARY 11: TOTAL READING

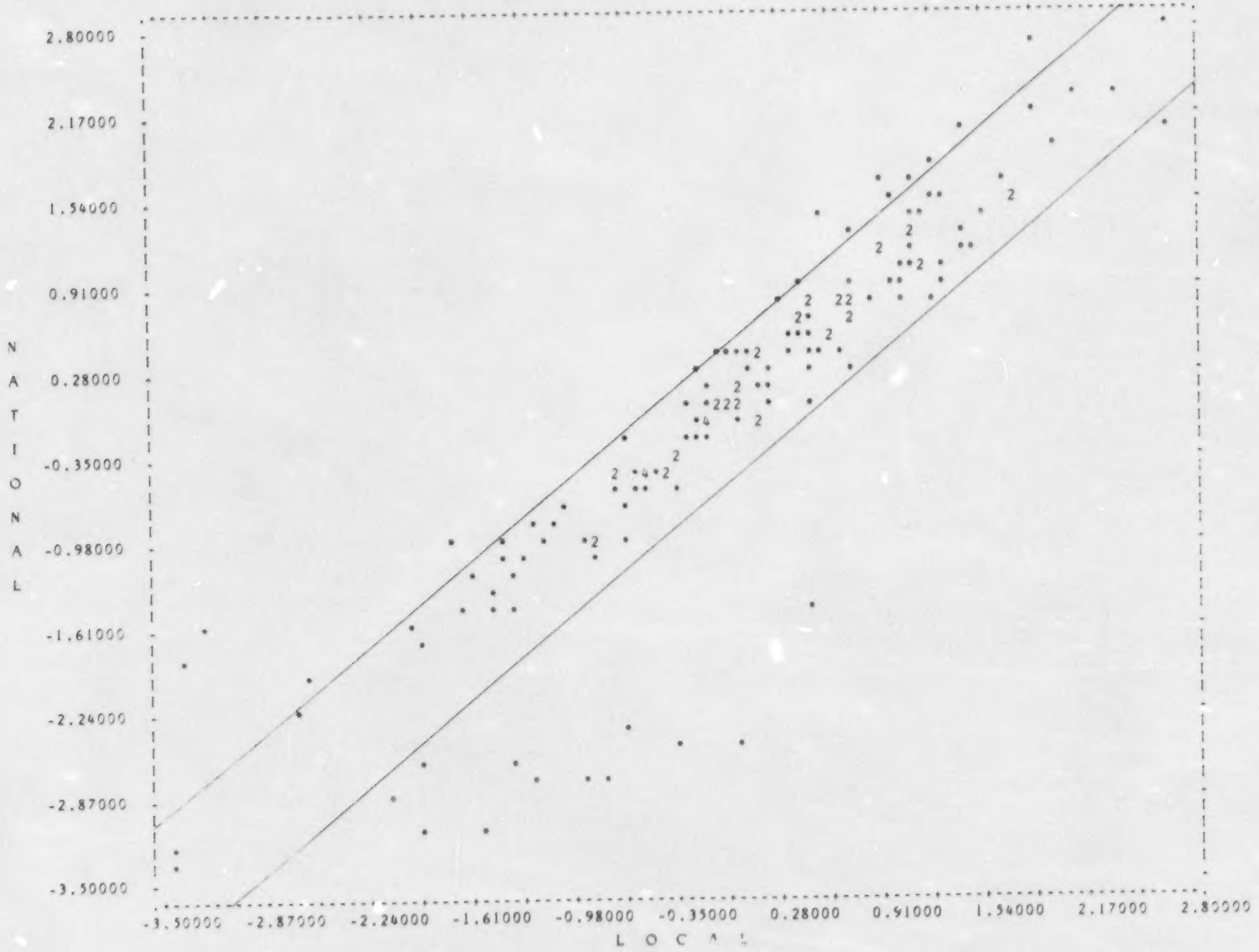


FIGURE 3

NATIONAL VS. LOCAL RASCH ITEM DIFFICULTIES FOR SAT/6 PRIMARY 1: TOTAL MATHEMATICS

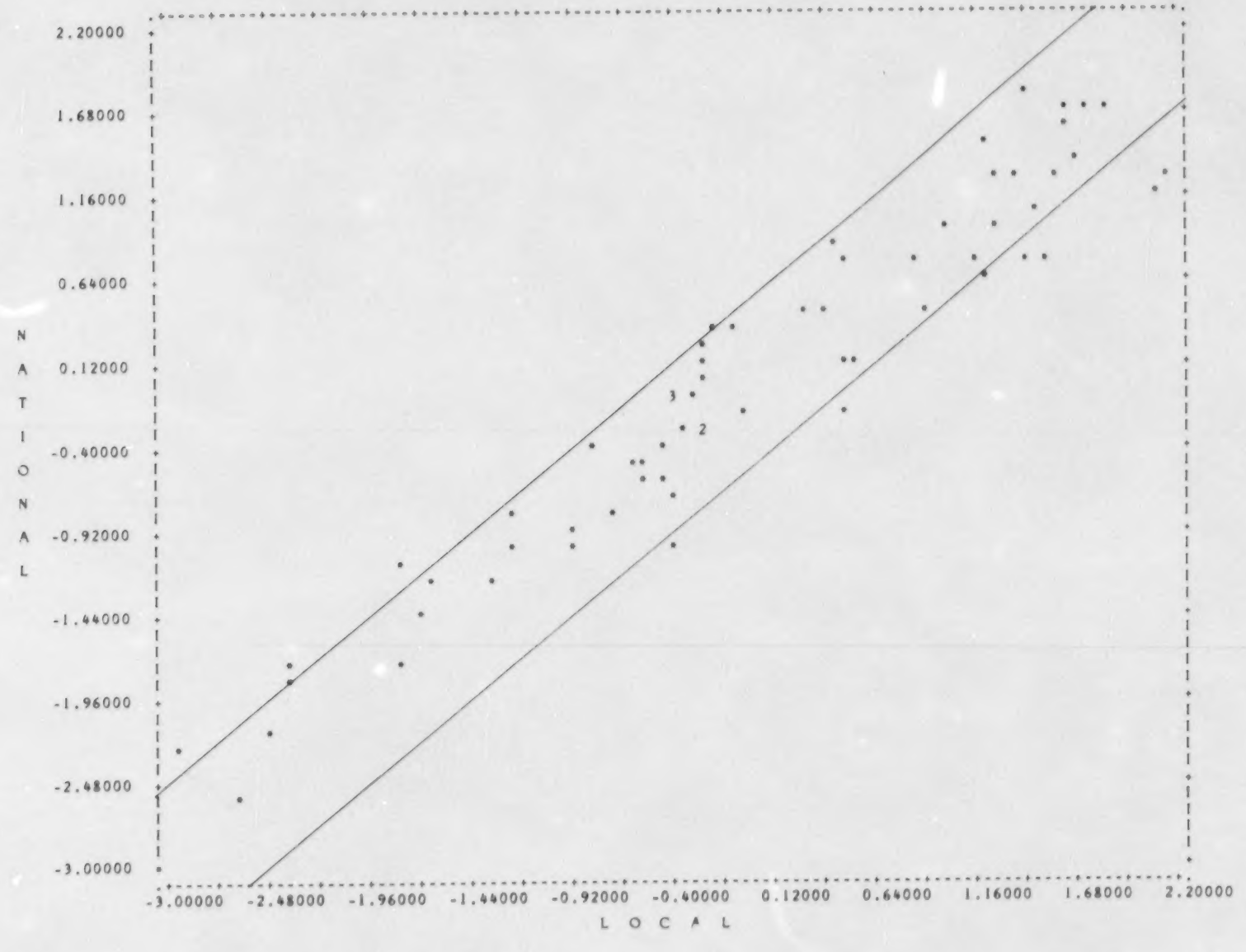
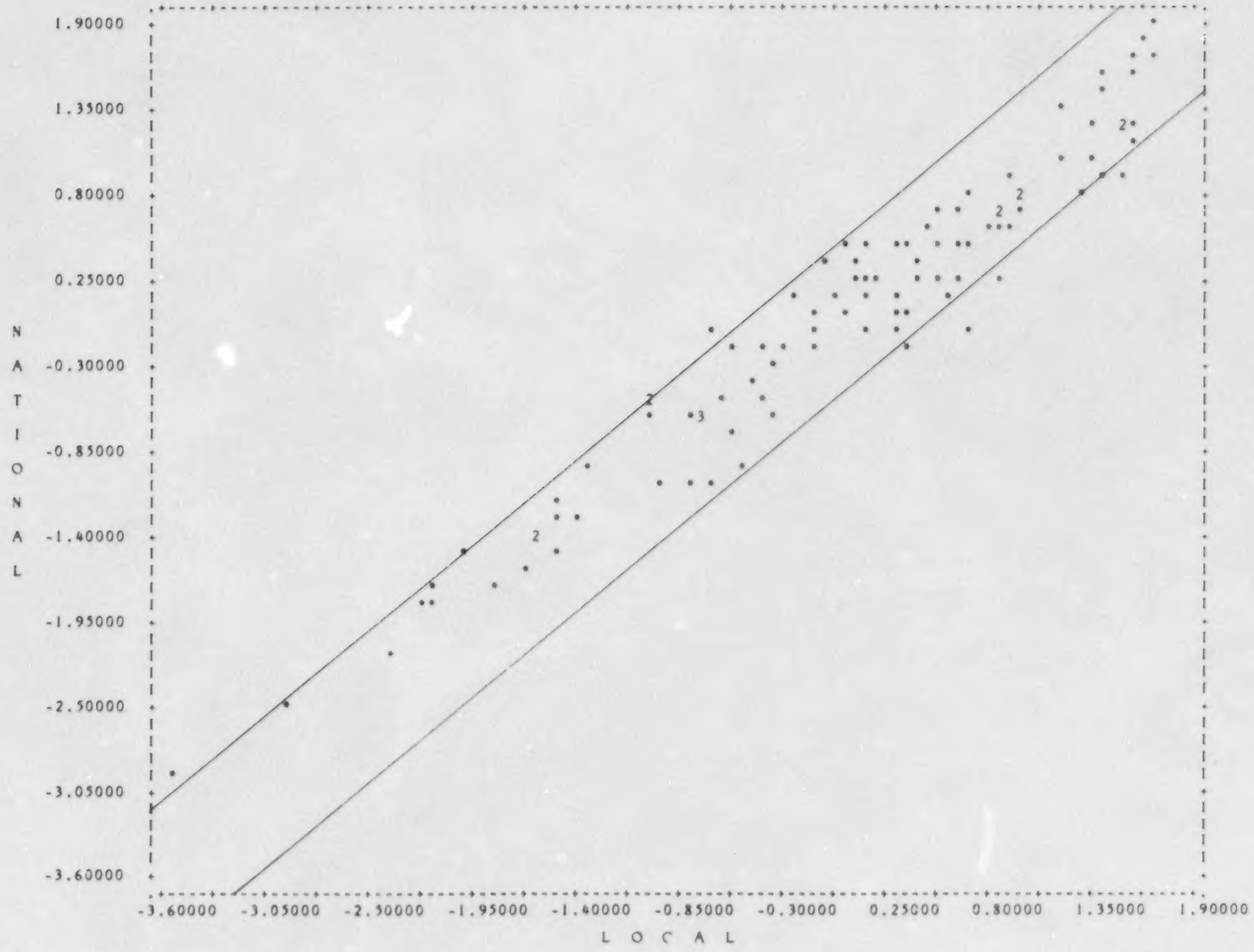


FIGURE 4

NATIONAL VS. LOCAL RASCH ITEM DIFFICULTIES FOR SAT/6 PRIMARY 11: TOTAL MATHEMATICS



SUPPLEMENTING STANDARDIZED ACHIEVEMENT TEST ITEMS FOR TESTING OF LOCAL OBJECTIVES

BERNARD G. JONES
GARY W. GRAMENZ

PALM BEACH COUNTY SCHOOLS

Presented at the

Annual Meeting of the

National Council on Measurement in Education

New York City

March, 1982

Supplementing Standardized Achievement Test Items for Testing of Local Objectives

Introduction

In an age of accountability, the current use of norm-referenced tests to evaluate the effectiveness of school programs is under attack by educators and others concerned with education. They argue that the educational objectives assessed by such tests often differ from what school systems are seeking to teach. Further, they maintain that the tests generally have been designed to sort students for various administrative purposes rather than to determine how well they have learned what is being taught. Many state education agencies have reacted to these criticisms by enacting legislation mandating objective-referenced minimum competency testing programs. Florida is one of these states.

In Florida, the State Student Assessment Testing Program provides objective-by-objective diagnostic reports which enable teachers to stress skill areas in which students perform poorly. Teachers and administrators have found this process to be very effective, and student performance on state minimum competency tests has improved steadily over the years. Following the lead of the state testing program, the Palm Beach County Schools has begun developing objective-referenced tests to provide diagnostic information on local curriculum objectives. This paper describes a method of constructing mathematics tests for grades 2-6 from a combination of locally-developed test items and standardized achievement test items.

Test Development

The first step in the developmental process was to determine the degree of match between SAT/7 mathematics items and local mathematics objectives. Dr. Bower spoke with you about this process. In some instances, one or more SAT/7 items tapped several on-grade-level objectives, while no item matches were found for other on-grade-level skills. Those items identified as measuring on-grade-level objectives comprise the SAT/7 subset of items to be used in the objective-referenced testing program.

Having matched SAT/7 mathematics items with local mathematics objectives, it was a simple matter to identify those objectives which were not measured by any Stanford item. The proportion of mathematics objectives not measured by SAT/7 items varied widely across grade levels (Table 1). For example, 47 percent of the mathematics objectives for grade three were not assessed by SAT/7 items, whereas only 30 percent of the second grade objectives were not measured.

Table 1

Number and Percent of Mathematics Objectives
Not Measured by SAT/7 Series

Grade Level	SAT/7 Level	Total No. Objectives	Objectives Not Measured	
			N	%
2	Pri. II (E)	46	14	30
3	Pri. III (E)	38	18	47
4	Int. I (E)	48	20	42
5	Int. II (E)	44	17	39
6	Int. II (F)	43	20	46

In order to maintain reasonable test lengths, weights ranging from 1 (least important) to 5 (most important) were assigned to objectives. Test items were selected to measure only objectives with assigned weights of 4 or 5, which reduced the number of objectives to be tested (Table 2). For example, although the local curriculum contains 46 second

Table 2

Number and Percent of Objectives Selected for Objective-Referenced Testing Program

Grade Level	Total No. Objectives	No. Objectives Selected	Percent Selected
2	46	36	78
3	38	22	58
4	48	28	58
5	44	31	70
6	43	33	77

grade mathematics objectives, only 36 were deemed most important (i.e., assigned weights of 4 or 5) for further progress in mathematics.

From the objective weighting and item matching procedures, the objectives to be measured and the SAT/7 items matching those objectives were determined. It was felt that, to obtain reliable measures, four items were needed to assess each objective. To measure student progress on objectives not measured by the Stanford series, or measured by less than four Stanford items, many additional items were required (Table 3).

Table 3

Number of Local Mathematics Items Needed
for Objective-Referenced Testing Program

Grade Level	Local
2	76
3	52
4	70
5	78
6	92

The number of local items needed to supplement the SAT/7 items for the objective-referenced testing program ranged from a low of 52 at grade three to a high of 92 at grade six.

Fortunately, another process was well underway that allowed us to provide these items. In 1979, the district had begun the development of a mathematics item bank for grades 1-6. Using the mathematics objectives that the district had developed in 1974 and refined in 1978, a test consultant firm generated item specifications which guided local item writers in constructing of items to measure the objectives. The items were reviewed by an outside consultant to determine the extent to which they were consonant with the item specifications. From a pool of approximately 900 items, 24 interlinking tests were constructed and administered to about 18,000 students in grades 4-6 in April of 1981. Items were assigned to the various forms and levels to allow us, using Rasch calibration techniques, to develop a mathematics item bank. From

the calibrated item bank, three forms (one per grade level) were constructed to be given as supplements to the SAT/7 mathematics tests to be administered in grades 4-6 in April of 1982.

The district plan also calls for expanding the bank downward to include mathematics items for grades 1-3. Fifteen field test forms (five per grade level) will be administered to students during the 1982 spring testing program. Calibration of these items will be completed by the summer of 1982. From the expanded bank, supplementary tests will be constructed and administered to students in grades 2-3 in April of 1983. (Additionally, the district has begun work on the development of a bank of communication skills items for grades 2-6.)

Reports

Two types of reports will be generated from the results of the objective-referenced testing program. The first is a diagnostic report on individual student mastery of each mathematics objective. (Mastery is defined as correctly answering three out of four items measuring an objective.)

The second type of report relies heavily on Rasch test equating methodology. A total score, based on student performance on all items across all objectives, will be computed. This score, expressed in Rasch scaled score units, will provide a yardstick with which to measure the year-to-year performance of individual students, as well as the overall degree of mastery of district mathematics objectives.

To generate this overall score, Rasch methodology will be used to calibrate items from each SAT/7 grade-level test and local supplementary test as one test. Since the local items are drawn from a Rasch calibrated item bank, a relationship will be established between the

SAT/7 mathematics items and the common scale of difficulty underlying the local bank items through the equating procedure. In essence, nationally standardized test items and locally-developed items will share a common metric and, hence, will constitute a common item pool.

As local objectives change over time, Stanford mathematics test items will be reexamined for curriculum match and supplementary tests will be revised to include items measuring new objectives and/or to delete items measuring objectives dropped from the local curriculum. The item bank concept is particularly useful in this regard. Since all subsets of items drawn from a calibrated bank are on a common scale, test scores from any revised form are automatically equated, thereby virtually eliminating the problem of test equating (Rentz and Bashaw, 1975 and 1977).

Summary

The combination of SAT/7 items and local supplementary items will be used to measure individual and aggregate student performance at one point in time and over time. Although our curriculum will change during the years we use the SAT/7, we have the flexibility to select any appropriate subsets of SAT/7 and supplementary test items to measure student performance on local objectives. The district can simply select different combinations of SAT/7 and local items for use in its objective-referenced testing program.

C/TEST/SSAT

REFERENCES

Rentz, R. R. and Bashaw, W. L. Equating reading tests with the Rasch model. University of Georgia Educational Research Laboratory, Athens, GA., 1975.

Rentz, R. R. and Bashaw, W. L. The national scale for reading: An application of the Rasch model. Journal of Educational Measurement, 1977, 14, 161-180.

APPENDIX

TWO-YEAR PROJECTION FOR CUSTOMIZATION OF STANFORD ACHIEVEMENT TEST

	SPRING 1982		SPRING 1983	
	Test(s) Administered	Scores Available*	Test(s) Administered	Scores Available
Grade 2	Primary 2 - E Supplementary Mathematics Test (field test form)	Traditional Norm-Referenced	Primary 2 - E Supplementary Mathematics Test	Traditional Norm-Referenced Curriculum-Referenced Norm (reading and mathematics) Objective-Referenced
Grade 3	Primary 3 - E Supplementary Mathematics Test (field test form)	Traditional Norm-Referenced	Primary 3 - E Supplementary Mathematics Test	Traditional Norm-Referenced Curriculum-Referenced Norm (reading and mathematics) Objective-Referenced
Grade 4	Intermediate 1 - E Supplementary Mathematics Test	Traditional Norm-Referenced Objective-Referenced	Intermediate 1 - E Supplementary Mathematics Test	Traditional Norm-Referenced Curriculum-Referenced Norm (reading and mathematics) Objective-Referenced
Grade 5	Intermediate 2 - E Supplementary Mathematics Test	Traditional Norm-Referenced	Intermediate 2 - E Supplementary Mathematics Test	Traditional Norm-Referenced Curriculum-Referenced Norm (reading and mathematics) Objective-Referenced
Grade 6	Intermediate 2 - F Supplementary Mathematics Test	Traditional Norm-Referenced Objective-Referenced	Intermediate 2 - F Supplementary Mathematics Test	Traditional Norm-Referenced Curriculum-Referenced Norm (reading and mathematics) Objective-Referenced
Grade 7	Advanced F	Traditional Norm-Referenced	Advanced F	Traditional Norm-Referenced
Grade 8	Advanced F	Traditional Norm-Referenced	Advanced F	Traditional Norm-Referenced
Grade 9	Advanced E	Traditional Norm-Referenced	Advanced E	Traditional Norm-Referenced

*Will be provided when national standardization is completed.

1/209(1A):9